Near-Optimal Estimation of Simultaneously Sparse and Low-Rank Matrices from Nested Linear Measurements

Sohail Bahmani,
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
sohail.bahmani@ece.gatech.edu
Justin Romberg
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
jrom@ece.gatech.edu

March 22, 2016

Abstract

In this paper we consider the problem of estimating simultaneously low-rank and row-wise sparse matrices from nested linear measurements where the linear operator consists of the product of a linear operator $\mathcal{W}$ and a matrix $\boldsymbol{\Psi}$. Leveraging the nested structure of the measurement operator, we propose a computationally efficient two-stage algorithm for estimating the simultaneously structured target matrix. Assuming that $\mathcal{W}$ is a restricted isometry for low-rank matrices and $\boldsymbol{\Psi}$ is a restricted isometry for row-wise sparse matrices, we establish an accuracy guarantee that holds uniformly for all sufficiently low-rank and row-wise sparse matrices with high probability. Furthermore, using standard tools from information theory, we establish a minimax lower bound for estimation of simultaneously low-rank and row-wise sparse matrices from linear measurements that need not be nested. The accuracy bounds established for the algorithm, that also serve as a minimax upper bound, differ from the derived minimax lower bound merely by a polylogarithmic factor of the dimensions. Therefore, the proposed algorithm is nearly minimax optimal. We also discuss some applications of the proposed observation model and evaluate our algorithm through numerical simulation.

1 Introduction

In this paper we study the problem of estimating a matrix $\boldsymbol{X}^\star$ that is simultaneously low-rank and row-wise sparse from the noisy linear measurements

$$\boldsymbol{y} = \mathcal{A}\left(\boldsymbol{X}^\star\right) + \boldsymbol{z}, \tag{1}$$

where $\mathcal{A}$ is a known linear operator and $\boldsymbol{z}$ models the additive error or noise. In particular, we consider this problem in the high-dimensional regime where the dimension of the measurements $\boldsymbol{y}$ can be much smaller than the ambient dimension of the target $\boldsymbol{X}^\star$. Estimation of low-rank matrices from the linear measurements (1) has been studied extensively in various settings (see, e.g., [14], [36], and [23]). However, in cases where the target matrix is characterized by two or more structures

simultaneously, the research regarding efficient and statistically optimal estimation is less developed. Ideally, different structures of the target matrix can be exploited to improve the sample complexity.

In this paper, assuming that the measurement operator $\mathcal{A}$ has a nested structure as described below, we address the problem of estimating the simultaneously low-rank and row-wise sparse $\boldsymbol{X}^\star$ from the linear measurements (1). Throughout the paper we may refer row-wise sparse matrices simply as sparse matrices unless explicitly stated otherwise. The nested operator $\mathcal{A}$ is the product of a linear operator and a matrix which we assume to be *restricted (near) isometries* for low-rank and row-wise sparse matrices, respectively. The inner matrix component in $\mathcal{A}$ compresses the columns of the target matrix without destroying its low-rank structure. This compression can be inverted because the target has sparse columns. The outer linear operator in $\mathcal{A}$ exploits the low-rank structure in the compressed matrix to compress the dimensions even more. We propose a computationally efficient method that relies on the nested structure of the measurement operator. We show that the proposed method is accurate uniformly for all simultaneously low-rank and row-wise sparse targets. Therefore, the obtained accuracy bound can be interpreted as an upper bound for the minimax rate as well. Furthermore, using standard information theoretic tools we establish a minimax lower bound that holds for any well-bounded linear operator $\mathcal{A}$, nested or otherwise. As will be seen below, the derived minimax bounds differ only by a polylogarithmic factor of the dimensions which confirm (near) optimality of the proposed estimation method.

1.1 Notation

Let us first set the notation used throughout the paper. Matrices and vectors are denoted by bold capital and small letters, respectively. The Hermitian adjoint of linear operators and matrices is denoted by a superscript asterisk such as $\mathcal{M}^*$ and $\boldsymbol{M}^*$. The set of positive integers less than or equal to n is denoted by $[n]$. For any $a \times b$ matrix $\boldsymbol{M}$ and a set $S \subseteq [b]$, the $a \times |S|$ restriction of $\boldsymbol{M}$ to the columns indexed by S is denoted by $\boldsymbol{M}_S$. In particular, $\boldsymbol{I}_{a,S}$ denotes the restriction of the $a \times a$ identity matrix to the columns indexed by S. The all-zero and all-one matrices of size $a \times b$ are denoted by $\boldsymbol{0}_{a\times b}$ and $\boldsymbol{1}_{a\times b}$, respectively. The Kronecker product of matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ is denoted by $\boldsymbol{M}_1 \otimes \boldsymbol{M}_2$. The orthogonal complement of any subspace V is denoted by $V^\perp$. The maximum and minimum of two numbers a and b are denoted by $a \vee b$ and $a \wedge b$, respectively. The notation $f = \mathsf{O}(g)$ is used when $f = cg$ for some absolute constant $c > 0$. We may also use $f \lesssim g$ (or $f \gtrsim g$) to denote $f \leq cg$ (or $f \geq cg$) for some absolute constant $c > 0$. The largest integer that is less than or equal to a real number a is denoted by $\lfloor a \rfloor$. Similarly, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. The function $d_{\mathsf{H}}(\cdot, \cdot)$ denotes the natural Hamming distance between a pair of objects that can be represented uniquely by finite binary sequences of the same length (e.g., binary vectors, binary matrices, subsets of a finite set, etc). For any matrix $\boldsymbol{M}$, the number of nonzero rows, the Frobenius norm, the nuclear norm, and the sum of the row-wise ℓ_2-norms (i.e., the $\ell_{1,2}$-norm) are denoted by $\|\boldsymbol{M}\|_{0,2}$, $\|\boldsymbol{M}\|_F$, $\|\boldsymbol{M}\|_*$, and $\|\boldsymbol{M}\|_{1,2}$, respectively.

1.2 Problem setup

We would like to estimate a $p_1 \times p_2$ matrix $\boldsymbol{X}^\star$ from noisy linear measurements given by (1) where the measurement operator $\mathcal{A} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is assumed to have the nested form

$$\mathcal{A}(\boldsymbol{X}) = \mathcal{W}(\boldsymbol{\Psi}\boldsymbol{X}), \tag{2}$$

with the matrix $\boldsymbol{\Psi} \in \mathbb{R}^{m\times p_1}$ and the linear operator $\mathcal{W} : \mathbb{R}^{m\times p_2} \to \mathbb{R}^n$ known. We also assume throughout the paper that $\boldsymbol{z}$ in (1) is a Gaussian noise with $\mathsf{N}\left(\mathbf{0}_{n\times 1}, \sigma^2 \boldsymbol{I}\right)$ distribution. The estimation problem is assumed to be in the high-dimensional regime (i.e., $n \ll p_1 p_2$). We consider the target matrices to have rank at most $r \ll p_1 \wedge p_2$ and at most $k \ll p_1$ nonzero rows, where clearly $r \leq k$. The set of all such matrices can be explicitly defined as

$$\mathbb{X}_{k,r} \coloneqq \left\{ \boldsymbol{X} \in \mathbb{R}^{p_1\times p_2} \mid \|\boldsymbol{X}\|_{0,2} \leq k,\ \mathsf{rank}\left(\boldsymbol{X}\right) \leq r \right\},$$

where $\|\boldsymbol{X}\|_{0,2}$ denotes the number of nonzero rows of $\boldsymbol{X}$. We denote the set of matrices with rank at most r by $\mathbb{X}_{\bullet,r}$ and the set of matrices with at most k nonzero rows by $\mathbb{X}_{k,\bullet}$. Furthermore, we assume that the rank parameter $r = \mathsf{rank}\left(\boldsymbol{X}^\star\right)$ and the noise variance σ^2 are known to the estimator.

The accuracy guarantees of our method can be treated as a minimax upper bound, if they hold uniformly for all target matrices in $\mathbb{X}_{k,r}$. To derive the desired uniform accuracy guarantees we assume that $\boldsymbol{\Psi}$ is a restricted isometry for sparse matrices and $\mathcal{W}$ is a restricted isometry for low-rank matrices which are common assumptions for compressive sensing and low-rank matrix estimation [5, 36, 13]. To be precise, for any integer $k \in [p_1]$, $\boldsymbol{\Psi}$ is a said to be a restricted isometry over $\mathbb{X}_{k,\bullet}$ if there exists a *restricted isometry constant* $\delta_{k,\bullet}\left(\boldsymbol{\Psi}\right) \in [0,1]$ such that

$$\forall \boldsymbol{X} \in \mathbb{X}_{k,\bullet} :\ \left(1 - \delta_{k,\bullet}\left(\boldsymbol{\Psi}\right)\right) \|\boldsymbol{X}\|_F^2 \leq \|\boldsymbol{\Psi}\boldsymbol{X}\|_F^2 \leq \left(1 + \delta_{k,\bullet}\left(\boldsymbol{\Psi}\right)\right) \|\boldsymbol{X}\|_F^2 .$$

Clearly, any restricted isometry for k-sparse vectors is a restricted isometry over $\mathbb{X}_{k,\bullet}$ and vice versa. Similarly, for all integers $r \in [m \wedge p_2]$, $\mathcal{W}$ is said to be the restricted isometry over $\mathbb{X}_{\bullet,r}$ if there exists a restricted isometry constant $\delta_{\bullet,r}\left(\mathcal{W}\right) \in [0,1]$ such that

$$\forall \boldsymbol{X} \in \mathbb{X}_{\bullet,r} :\ \left(1 - \delta_{\bullet,r}\left(\mathcal{W}\right)\right) \|\boldsymbol{X}\|_F^2 \leq \|\mathcal{W}\left(\boldsymbol{X}\right)\|_2^2 \leq \left(1 + \delta_{\bullet,r}\left(\mathcal{W}\right)\right) \|\boldsymbol{X}\|_F^2 .$$

The goal of our minimax approach is to characterize the tolerance level τ in terms of the parameters of the estimation problem, such that the "minimax failure probability" given by

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r}} \mathsf{P}\left(\left\| \widehat{\boldsymbol{X}} - \boldsymbol{X}^\star \right\|_F \geq \tau \right)$$

is sufficiently small. In words, τ would be the worst-case accuracy of the best estimator that holds with high probability. As will be seen in Section 2, we use the restricted isometry properties mentioned above to provide a uniform accuracy guarantee for our proposed method which also serves as a minimax upper bound. To establish the minimax lower bound, we employ some of the standard information theoretic tools that help to determine non-trivial values of τ for which the above minimax failure probability is significant, e.g., more than $\frac{1}{2}$. Derivation of the minimax lower bound does not depend on the restricted isometry assumptions and we merely assume that

$$\forall \boldsymbol{X} \in \mathbb{X}_{k,r} :\ \|\mathcal{A}\left(\boldsymbol{X}\right)\|_2^2 \leq \gamma_{k,r} \|\boldsymbol{X}\|_F^2 , \tag{3}$$

for some constant $\gamma_{k,r} > 0$. Of course, if $\mathcal{A}$ is defined by (2) and we happen to have the restricted isometries $\boldsymbol{\Psi}$ and $\mathcal{W}$ as above, clearly there exists a constant $\gamma_{k,r}$ such that $\gamma_{k,r} \leq \left(1 + \delta_{\bullet,r}\left(\mathcal{W}\right)\right)\left(1 + \delta_{k,\bullet}\left(\boldsymbol{\Psi}\right)\right)$.

1.3 Contributions

- **Minimax upper bound:**

  Under the observation model described by (1) and (2) we propose a method that

With $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$ and under the assumption that $\mathcal{W}$ and $\boldsymbol{\Psi}$ are restricted isometries over rank-$4r$ matrices (i.e., $\mathbb{X}_{\bullet,4r}$) and row-wise $2k$-sparse matrices (i.e., $\mathbb{X}_{2k,\bullet}$), respectively, produces some estimate $\widehat{\boldsymbol{X}}$ of $\boldsymbol{X}^\star$ such that

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \lesssim \sigma\sqrt{r\left(m \vee p_2\right)},$$

holds uniformly for all $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$.

In particular, we show that if $\boldsymbol{\Psi}$ and $\mathcal{W}$ are Gaussian with properly scaled iid entries then

With $m = \mathsf{O}\left(k \log \frac{p_1}{k}\right)$ and $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$ the desired restricted isometry properties hold and the produced estimate obeys

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \lesssim \sigma\sqrt{r\left(k \log \frac{p_1}{k} \vee p_2\right)},$$

for all $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$.

Because the above accuracy results hold uniformly for all targets in $\mathbb{X}_{k,r}$, they can also be viewed as a minimax upper bound.

- **Minimax lower bound:**

  We also establish a minimax lower bound merely under the assumption that the linear measurement operator $\mathcal{A}$ satisfies (3) whether it takes the nested form (2) or not. Namely, we show that

  If $\mathcal{A}$ obeys (3) then for a sufficiently small absolute constant $c > 0$ and for any estimator $\widehat{\boldsymbol{X}}$ there exists a target $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$ such that

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq c\sigma\sqrt{\frac{k \log \frac{p_1}{k} + r\left(k \vee p_2\right)}{\gamma_{k,r}}}$$

  holds with probability at least $\frac{1}{2}$.

With $\gamma_{k,r} = \mathsf{O}\left(1\right)$, the obtained minimax lower and upper bounds differ only by some logarithmic factors, which shows that by exploiting the nested form of the measurements the proposed estimator achieves a near optimal minimax rate. For comparison, if we only considered the low-rank structure, then for any estimator the estimation error in Frobenius norm would have been bounded by $\mathsf{O}\left(\sigma\sqrt{r\left(p_1 \vee p_2\right)}\right)$ from below (cf. [37] and [13, Theorem 2.5]). Similarly, considering only the sparsity as the structure of $\boldsymbol{X}^\star$, the accuracy lower bounds for sparse estimation (e.g.,[35, Theorem 1]) suggest the lower bound $\mathsf{O}\left(\sigma\sqrt{p_2 k \log \frac{p_1}{k}}\right)$ for the estimation error. Therefore, with the natural assumption that $p_1 \gg p_2$, these lower bounds indicate that exploiting the low-rank and sparse structures simultaneously has reduced the estimation error dramatically. Furthermore, in the noise-free scenario, our results suggest that we can recover $\boldsymbol{X}^\star$ exactly from $\mathsf{O}\left(r\left(k \log \frac{p_1}{k} \vee p_2\right)\right)$ measurements. Therefore, the nested structure of the measurements and the proposed estimation procedure have a critical role in achieving the mentioned sample complexity which cannot be achieved by minimization of any convex combination of the nuclear norm and the $\ell_{1,2}$-norm for many common types of linear measurements [34].

1.4 Related work

Many important problems in statistics and machine learning such as Sparse Principal Component Analysis (SPCA) can be formulated as estimation of simultaneously low-rank and sparse matrices. These estimation problems are significantly more challenging than estimation of matrices that are either low-rank or sparse. In *sparse spiked covariance* estimation [22] and more generally SPCA, given a relatively small number of independent samples of a mean-zero multivariate Gaussian, the primary goal is to estimate the principal eigenvectors of the associated covariance matrix that are assumed to be sparse. Many convex and non-convex algorithms have been proposed for these problems including but not limited to [17, 22, 1, 11, 43, 44, 10]. Each of these algorithms provides a different trade-off between statistical optimality and computational complexity. However, as shown in [7], there may be an inevitable gap between the optimal statistical accuracy and the statistical accuracy that polynomial-time algorithms can achieve in detection of sparse principal components. Since SPCA can be viewed as a primitive for solving linear regression with simultaneously low-rank and sparse target matrices, the result of [7] can also indicate the difficulty of the latter problem.

The problem of denoising low-rank and row-wise sparse matrices is also considered in [9]. The iterative algorithm proposed in [9] is basically an adaptation of the *power iteration* where the factors are sparsified through thresholding. It is shown, that under mild conditions, this iterative method can be initialized appropriately and then is shown to be nearly minimax optimal. A more related work to our problem is multiple linear regression considered in [28] where the goal is to estimate $\boldsymbol{X}^\star$ that is low-rank and row-wise sparse from the measurements $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}^\star + \boldsymbol{Z}$ with $\boldsymbol{A}$ and $\boldsymbol{Z}$ being the design and the noise matrices, respectively. The algorithm of [28] uses the low-rank structure of $\boldsymbol{X}^\star$ to construct an initial estimate $\boldsymbol{V}$ of the right singular vectors of $\boldsymbol{X}^\star$. Using this estimate, the algorithm then solves a least squares with a convex regularization for row-wise sparsity to estimate $\boldsymbol{B}$, the projection of $\boldsymbol{X}^\star$ onto the estimated singular vectors. Then $\boldsymbol{V}$ is updated to be the right singular vectors of $\boldsymbol{Y}$ after projecting its range onto the range of $\boldsymbol{B}$. Finally, the algorithm updates $\boldsymbol{B}$ by repeating the mentioned convex optimization and outputs $\boldsymbol{B}\boldsymbol{V}^\mathsf{T}$ as the estimate of $\boldsymbol{X}^\star$ which is shown to be nearly minimax optimal.

Furthermore, an efficient alternating minimization method for estimation of rank-one and sparse matrices from linear measurements is proposed in [26]. The algorithm is shown to be accurate and nearly optimal in terms of sample complexity, provided that the factors of the target rank-one matrix have relatively dominant spikes and the noise level is moderate. Estimation of simultaneously structured matrices from compressive linear measurements is also studied in [34]. It is shown in [34] that using convex proxies for each of the assumed structures independently would result in a suboptimal sample complexity. Specifically, it is shown [34] that minimization of any mixture of the $\ell_{1,2}$-norm and the nuclear norm fails to recover the true signal if there are less than $\mathsf{O}\left(kp_2 \wedge r\left(p_1 + p_2\right)\right)$ measurements. Therefore, our results as stated above show a significant improvement over the naïve convex relaxation method studied in [34].

2 Main Results

2.1 Two-stage estimator

We propose the following two-stage method for estimation of simultaneously low-rank and row-wise sparse matrices from measurements given by (1) when $\mathcal{A}$ has the nested structure described by (2). The nuclear norm and the sum of row-wise ℓ_2-norm are denoted by $\|\cdot\|_*$ and $\|\cdot\|_{1,2,}$, respectively.

1. Low-rank estimation stage:

$$\widehat{\boldsymbol{B}} \in \underset{\boldsymbol{B}}{\operatorname{argmin}} \|\boldsymbol{B}\|_* \quad \text{subject to } \|\mathcal{W}(\boldsymbol{B}) - \boldsymbol{y}\|_2 \leq \sigma\sqrt{n + c_1 r\,(m \vee p_2)} \tag{4}$$

2. Sparse estimation stage:

$$\widehat{\boldsymbol{X}} \in \underset{\boldsymbol{X}}{\operatorname{argmin}} \|\boldsymbol{X}\|_{1,2} \quad \text{subject to } \left\|\boldsymbol{\Psi}\boldsymbol{X} - \widehat{\boldsymbol{B}}\right\|_F \leq c_2\sigma\sqrt{r\,(m \vee p_2)} \tag{5}$$

With appropriately chosen constants c_1 and c_2 each stage is a convex program for which many efficient solvers exist.

Variations of the estimator: The specific form of the optimization that expresses the low-rank estimation and the sparse estimation stages is not critically important. It is only necessary to produce a solution that is sufficiently accurate in each stage. For example, with the appropriate regularization parameter $\lambda > 0$, we could have used the regularized analog of (4) given by

$$\widehat{\boldsymbol{B}} \in \underset{\boldsymbol{B}}{\operatorname{argmin}} \frac{1}{2}\|\mathcal{W}(\boldsymbol{B}) - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{B}\|_*$$

which also enjoys the desired accuracy guarantees [13]. Similarly, the sparse estimation stage can be performed by the regularized form of (5). Furthermore, because we are assuming that $\boldsymbol{\Psi}$ and $\mathcal{W}$ are restricted isometries, the convex programs considered in (4) and (5) could be replaced by non-convex greedy algorithms such as the *iterative hard thresholding* (IHT) [8] that achieve the same accuracy usually at a lower computational cost. These methods, however, usually require tighter bounds on the restricted isometry constants.

Post-processing: The solutions obtained by (4) and (5) generally are not low-rank or sparse. To enforce the desired structures, after each stage of the estimator we can simply project the estimator onto the set of low-rank and/or row-wise sparse matrices. It is straightforward to show that these post-processing steps will not change the derived error bounds beyond a constant factor. Furthermore, we can treat range of the best rank-r approximation to $\widehat{\boldsymbol{B}}$ as an estimate for the range of $\boldsymbol{\Psi}\boldsymbol{X}^\star$ and pass it to a sparse estimation stage similar to (5), but with an optimization variable that has r columns rather than p_2. Therefore, we can significantly reduce the computational cost of the second stage of the estimator. However, to analyze the performance of this modified estimator it is necessary to convert the range estimation to the actual estimation error which we do not pursue in this paper.

2.2 Accuracy of the estimator and the minimax upper bound

In this section we state our result on the statistical accuracy of the two-stage estimator described by (4) and (5). This accuracy guarantee can be viewed as a minimax upper bound as well.

For $\boldsymbol{z} \sim \mathsf{N}\left(\boldsymbol{0}_{n\times 1}, \sigma^2 \boldsymbol{I}\right)$, tail bounds for chi-squared random variables [25, Lemma 1] guarantee that for any $\nu > 0$ we have

$$\sigma^2 (n - \nu) \leq \|\boldsymbol{z}\|_2^2 \leq \sigma^2 (n + \nu)$$

with probability at least $1 - 2\exp\left(-\left(\frac{\nu}{4} \wedge \frac{\nu^2}{16n}\right)\right)$. Therefore, assuming that $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$ and choosing $\nu = c_1 r\left(m \vee p_2\right)$ for a sufficiently large absolute constant $c_1 > 0$ we have

$$\sigma^2 \left(n - c_1 r\left(m \vee p_2\right)\right) \leq \|\boldsymbol{z}\|_2^2 \leq \sigma^2 \left(n + c_1 r\left(m \vee p_2\right)\right), \tag{6}$$

with high probability. Therefore, we can guarantee that the matrix $\boldsymbol{B}^\star = \boldsymbol{\Psi}\boldsymbol{X}^\star$ is in the feasible set of (4) with high probability. Consequently, we can invoke Lemma 2 and guarantee that $\boldsymbol{X}^\star$ is in the feasible set of (5) for appropriate choice of the constant $c_2 > 0$. Of course, the constants c_1 and c_2 depend on the restricted isometry constants of $\mathcal{W}$ and $\boldsymbol{\Psi}$.

Theorem 1 (minimax upper bound)**.** *Let $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$. Furthermore, suppose that $\boldsymbol{\Psi}$ and $\mathcal{W}$ have sufficiently small restricted isometry constants $\delta_{\bullet, 2k}\left(\boldsymbol{\Psi}\right)$ and $\delta_{4r, \bullet}\left(\mathcal{W}\right)$, respectively. Then, for appropriately chosen constants c_1 and c_2, there exists an absolute constant $C > 0$ depending on c_1 and c_2 such that the estimate produced using (4) and (5) obeys*

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \leq C\sigma\sqrt{r\left(m \vee p_2\right)}$$

for all $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$ with high probability.

The proof of Theorem 1 is straightforward and provided in the appendix. If $\boldsymbol{\Psi}$ and $\mathcal{W}$ are drawn from certain ensembles of random matrices or operators (e.g., Gaussian, Rademacher, partial Fourier, partial random circulant, etc.), then with

$$m = \mathsf{O}\left(k \operatorname{polylog}\left(p_1, k\right)\right) \qquad \text{and} \qquad n = \mathsf{O}\left(r\left(k \vee p_2\right) \operatorname{polylog}\left(p_1, p_2, k, r\right)\right),$$

where $\operatorname{polylog}()$ denotes a polylogarithmic factor of its argument, we can guarantee the desired restricted isometry properties with high probability. Therefore, an immediate consequence of Theorem 1 in these scenarios is that we can guarantee estimation error of the order $\sqrt{r\left(k \vee p_2\right) \operatorname{polylog}\left(p_1, p_2, k, r\right)}$. To have a concrete example, the case of Gaussian operators is addressed by the following corollary.

Corollary 1. *Suppose that $\boldsymbol{\Psi}$ has iid $\mathsf{N}\left(0, \frac{1}{m}\right)$ entries. Furthermore, suppose that $\mathcal{W}$ is a Gaussian operator that simply takes the inner product of its argument with n independent Gaussian matrices each populated with iid $\mathsf{N}\left(0, \frac{1}{n}\right)$ entries. If $m = C_1 k \log\frac{p_1}{k}$ and $n = C_2 r\left(m \vee p_2\right)$ for sufficiently large absolute constants C_1 and C_2, then there exists an absolute constant $C > 0$ such that the estimate $\widehat{\boldsymbol{X}}$ obtained using (4) and (5) satisfies*

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \leq C\sigma\sqrt{r\left(k\log\frac{p_1}{k} \vee p_2\right)},$$

for all $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$ with high probability.

Proof. With $m = \mathsf{O}\left(k \log \frac{p_1}{k}\right)$ it can be shown through the standard covering argument and the union bound that, with high probability, the considered $\boldsymbol{\Psi}$ would be a restricted isometry over $2k$-sparse vectors in $\mathbb{R}^{p_1}$ [5]. Clearly, in this case $\boldsymbol{\Psi}$ is also a restricted isometry over $\mathbb{X}_{2k,\bullet}$ as it preserves the ℓ_2-norm of the columns of any matrix in $\mathbb{X}_{2k,\bullet}$. Similarly, $\mathcal{W}$ would be a restricted isometry over rank-r matrices in $\mathbb{R}^{m \times p_2}$ if $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$ [13, Theorem 2.3]. Then, the desired accuracy bound follows immediately from Theorem 1. □

Weaker assumptions and non-uniform guarantees: The assumptions that the linear operator $\mathcal{W}$ and the matrix $\boldsymbol{\Psi}$ are restricted isometries are primarily used to establish the minimax bounds that are valid uniformly for all target matrices $\boldsymbol{X}^\star \in \mathbb{X}_{k,r}$. However, these assumptions are not generally needed if the goal is merely to show accuracy of the proposed two-stage method for any particular instance of the problem. For example, for certain random operators $\mathcal{W}$ and matrices $\boldsymbol{\Psi}$ that cannot be restricted isometries, the accuracy of (1) and (4) can be shown by construction of a *dual certificate* through the *golfing scheme* [20]. In these regimes, however, the robustness to noise is often weaker.

Low-rank and column-sparse matrices: If the target low-rank matrices are column-sparse rather than row-wise sparse our results still apply with minor adjustments. if the columns of a low-rank matrix are k-sparse, but not necessarily supported on the same rows, the restricted isometry of $\boldsymbol{\Psi}$ still holds. Therefore, we can follow a similar algorithm in which the $\ell_{1,2}$ norm in the second stage is replaced by another norm such as the maximum column-wise ℓ_1 norm. The lower bound still holds as well since row-wise sparse matrices are a special instances of column-sparse matrices.

2.3 The minimax lower bound

The following theorem provides a minimax lower bound for the probability of estimation failure over $\mathbb{X}_{k,r}$.

Theorem 2 (minimax lower bound)**.** *Suppose that $\mathcal{A}$ obeys (3). Let $\widehat{\boldsymbol{X}}$ denote any estimator of $\boldsymbol{X}^\star$ based on the measurements of the form (1). Then, there exists a sufficiently small absolute constant $c > 0$ such that with a probability more than one half the estimation error over $\mathbb{X}_{k,r}$ exceeds $c\sigma\sqrt{\frac{k \log \frac{p_1}{k} + r(k \vee p_2)}{\gamma_{k,r}}}$. Namely, we have the minimax lower bound*

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq c\sigma\sqrt{\frac{k \log \frac{p_1}{k} + r\left(k \vee p_2\right)}{\gamma_{k,r}}}\right) > \frac{1}{2}.$$

In words, Theorem 2 shows that if $\gamma_{k,r} = \mathsf{O}(1)$, then the error of any estimator of the matrices in $\mathbb{X}_{k,r}$ cannot be uniformly better than $\mathsf{O}\left(\sigma\sqrt{k \log \frac{p_1}{k} + r\left(k \vee p_2\right)}\right)$ with high probability. For instance, this result implies that the bound established in Corollary 1 is near-optimal as the error bound is within $\log\left(\frac{p_1}{k}\right)$ factor of the lower bound. We would like to emphasize that, we did not assume the nested structure (2) for $\mathcal{A}$ to prove the lower bound. Therefore, the lower bound applies to any linear operator $\mathcal{A}$ obeying (3) regardless of whether it is of the form (2) or not.

3 Applications and Extensions

3.1 Blind deconvolution with randomly coded masks

The linear measurements of the form (2) appear in some interesting and important applications in computational imaging. For example, (2) can describe the measurements in the lifted formulation of blind deconvolution problems that arise in imaging systems with randomly coded masks [40, 3]. In these problems rank-one target matrix is $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{h}^*$ with $\boldsymbol{u}$ and $\boldsymbol{h}$ being the target image and the Discrete Fourier Transform (DFT) of the unknown blurring kernel, respectively. The blind deconvolution can be solved effectively, by estimating the matrix $\boldsymbol{X}$ from the linear measurements of the form

$$\mathcal{A}(\boldsymbol{X}) = \boldsymbol{\Phi}^* (\boldsymbol{X} \circ \boldsymbol{F}).$$

where $\boldsymbol{F}$ denotes the DFT matrix, $\circ$ denotes the Hadamard product, and $\boldsymbol{\Phi}$ is a matrix whose columns model the random masks that modulate the image $\boldsymbol{u}$. We assume that the image $\boldsymbol{u}$ is sparse with respect to an orthonormal basis that is *incoherent* with the canonical basis. This assumption is realistic because, if necessary, spectral modulation can always create an effective sparsifying basis for the target image that has the desired incoherence condition. To simplify the exposition, let us assume that the image is sparse in the DFT basis and we have $\boldsymbol{u} = \boldsymbol{F}\tilde{\boldsymbol{u}}$ with $\widetilde{\boldsymbol{u}}$ being a k-sparse vector. Therefore, instead of $\boldsymbol{X}$ one may consider $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{u}}\boldsymbol{h}^*$ as the target rank-one matrix which is also row-wise sparse. Now if the random masks $\boldsymbol{\Phi}$ are designed to suppress all but a small number of randomly chosen pixels indexed by Ω, then the measurement model reduces to

$$\widetilde{\mathcal{A}}\left(\widetilde{\boldsymbol{X}}\right) = \boldsymbol{\Phi}_\Omega^* \left(\boldsymbol{F}_\Omega \widetilde{\boldsymbol{X}} \circ \boldsymbol{F}_\Omega\right),$$

where the subscript Ω denotes restriction to the rows indexed by Ω. It is now clear that with $\boldsymbol{\Psi} = \boldsymbol{F}_\Omega$ and $\mathcal{W} : \boldsymbol{B} \mapsto \mathsf{vec}\left(\boldsymbol{\Phi}_\Omega^* (\boldsymbol{B} \circ \boldsymbol{F}_\Omega)\right)$, the above equation is a special case of (2) up to a trivial vectorization.

3.2 Low-rank and doubly-sparse matrices

In this paper, we chose to state the main theoretical results only for the low-rank and row-wise sparse model mentioned in Section 1.2. However, these results can also be easily extended to the case of low-rank and doubly sparse matrices, where the target matrix is sparse both row-wise and column-wise. These kinds of matrices occur, for instance, in compressive phase retrieval, elaborated on in the following subsection, and *covariance sketching* [4]. The appropriate nested form of the measurement operator for these problems is

$$\mathcal{A}(\boldsymbol{X}) = \mathcal{W}(\boldsymbol{\Psi}_1 \boldsymbol{X} \boldsymbol{\Psi}_2^*), \tag{7}$$

with $\mathcal{W}$, $\boldsymbol{\Psi}_1$, and $\boldsymbol{\Psi}_2$ given. The only modification needed for the estimator would be to use $\boldsymbol{\Psi}_1$ and perform the first sparse estimation stage for row-wise sparsity as before, and then use $\boldsymbol{\Psi}_2$ to perform a second sparse estimation stage for column-wise sparsity.

3.3 Compressive phase retrieval

An interesting special case of the extension described above by (7) is when $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2$ and the linear operator $\mathcal{W}$ measures inner product of its argument and some rank-one matrices $\boldsymbol{w}_i\boldsymbol{w}_i^*$ (i.e., $\mathcal{W} : \boldsymbol{B} \mapsto [\langle \boldsymbol{w}_i\boldsymbol{w}_i^*, \boldsymbol{B}\rangle]_{i=1}^n$). In particular, this model can be used in Compressive Phase Retrieval (CPR) [31, 39] where the target matrix $\boldsymbol{X}^\star = \boldsymbol{x}\boldsymbol{x}^*$ with a sparse $\boldsymbol{x} \in \mathbb{C}^p$ is estimated from measurements $\left[|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2\right]_{i=1}^n$. Therefore, the measurement operator in CPR can be written as $\mathcal{A} : \boldsymbol{X} \mapsto [\langle \boldsymbol{a}_i\boldsymbol{a}_i^*, \boldsymbol{X}\rangle]_{i=1}^n$. The CPR problem is posed as non-convex optimization problems in [31, 39, 38] where the sparsity of the solution is minimized subject to a constraint on the quartic prediction error or vice versa. While certain local convergence guarantees are established for the GESPAR algorithm [38], global convergence and statistical accuracy of the proposed algorithm remains unknown. In [32], another non-convex approach based on alternating minimization is proposed for the standard Phase Retrieval (PR) as well as the CPR. With the particular initialization proposed in [32], the alternating minimization method is shown to converge linearly in the noise-free PR and CPR problems. However, this convergence rate holds for CPR when the number of measurements grows quadratically in k. For large-scale problems this alternating minimization method would still be a favorite choice among the competing algorithms as it is computationally less demanding. Furthermore, assuming iid Gaussian measurement vectors $\boldsymbol{a}_i$, [33] and [27] consider a convex relaxation to the lifted CPR problem formulated as

$$\begin{aligned} \underset{\boldsymbol{X}\succcurlyeq\boldsymbol{0}}{\text{argmin}} \quad & \mathsf{trace}\,(\boldsymbol{X}) + \lambda \left\|\boldsymbol{X}\right\|_1 \\ \text{subject to} \quad & \mathcal{A}\,(\boldsymbol{X}) = \boldsymbol{y}, \end{aligned}$$

for some parameter $\lambda \geq 0$. For the considered type of measurements, [27] shows that the above convex program can recover the target sparse and rank-one matrix $\boldsymbol{X}^\star$ if the number of measurements depend quadratically on k, the number of the nonzeros of the signal $\boldsymbol{x}$. The same result is shown in [16] for sub-Gaussian measurements and through a simpler derivation by "debiasing" the measurements and showing the obtained operator obeys an ℓ_1/ℓ_2 variant of the restricted isometry property. Furthermore, [27] demonstrates that the quadratic dependence on k cannot be improved fundamentally by varying the coefficient λ which is in agreement with the results of [34].

As mentioned above, if we choose $\boldsymbol{a}_i = \boldsymbol{\Psi}^*\boldsymbol{w}_i$ for some compressive sensing matrix $\boldsymbol{\Psi}$, and random vectors $\{\boldsymbol{w}_i\}_{i=1}^n$, then the measurement operator takes the nested form $\mathcal{A} : \boldsymbol{X} \mapsto [\langle \boldsymbol{w}_i\boldsymbol{w}_i^*, \boldsymbol{\Psi}\boldsymbol{X}\boldsymbol{\Psi}^*\rangle]_{i=1}^n$. It is then straightforward to apply the framework developed in this paper and show that $\boldsymbol{X}^\star$ can be reconstructed accurately and efficiently from $n \ll p$ measurements obtained by $\mathcal{A}$ through the two-stage recovery

$$\begin{aligned} \widehat{\boldsymbol{B}} \in \underset{\boldsymbol{B}\succcurlyeq\boldsymbol{0}}{\text{argmin}}\ & \mathsf{trace}\,(\boldsymbol{B}) \qquad (8) \\ \text{subject to}\ & \sum_{i=1}^n \left(\boldsymbol{w}_i^*\boldsymbol{B}\boldsymbol{w}_i - y_i\right)^2 \leq \varepsilon^2 \\ \widehat{\boldsymbol{X}} \in \underset{\boldsymbol{X}}{\text{argmin}}\ & \left\|\boldsymbol{X}\right\|_1 \qquad (9) \\ \text{subject to}\ & \left\|\boldsymbol{\Psi}\boldsymbol{X}\boldsymbol{\Psi}^* - \widehat{\boldsymbol{B}}\right\|_F \leq \frac{C\varepsilon}{\sqrt{n}}, \end{aligned}$$

where $\varepsilon \geq \left\|\boldsymbol{z}\right\|_2$ is a bound on the noise, $\left\|\cdot\right\|_1$ denotes the ℓ_1 norm, and C is an absolute constant. The

precise guarantees for effectiveness of nested measurements in CPR are established independently in [21], [46], and [2]. In particular, the following is shown in [2].

Theorem 3 ([2, Corollary 1])**.** *Let $\boldsymbol{\Psi} \in \mathbb{R}^{m\times p}$ be a matrix with independent $\mathsf{N}(0, \frac{1}{m})$ entries, and for $i = 1, 2, \dots, n$ let $\boldsymbol{w}_i \in \mathbb{R}^m$ be independent copies of a vector with i.i.d. standard Gaussian entries. For an arbitrary k-sparse signal $\boldsymbol{x} \in \mathbb{R}^p$ we observe measurements of the form $y_i = \langle \boldsymbol{\Psi}^\mathsf{T} \boldsymbol{w}_i, \boldsymbol{x}\rangle^2 + z_i$. If $m \geq c_1 k \log \frac{p}{k}$ and $n \geq c_2 m$ for sufficiently large absolute constant c_1 and c_2, then the two-stage recovery through* (8) *and* (9) *produces an estimate $\widehat{\boldsymbol{X}}$ that obeys*

$$\left\| \widehat{\boldsymbol{X}} - \boldsymbol{x}\boldsymbol{x}^\mathsf{T} \right\|_F \leq \frac{C'\varepsilon}{\sqrt{n}},$$

with high probability for some absolute constant $C' > 0$.

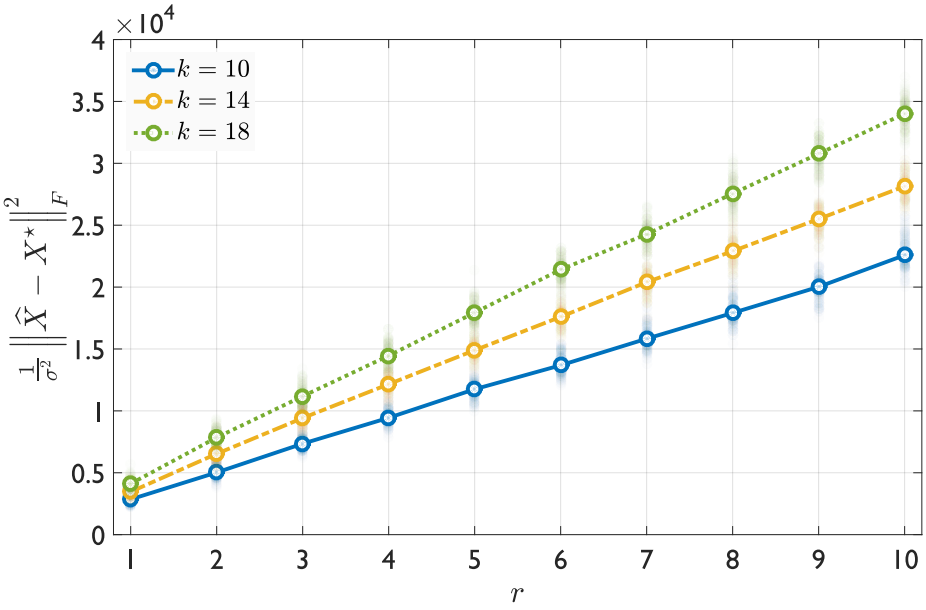
The fact that the intermediate operator $\mathcal{W} : \boldsymbol{B} \mapsto [\langle \boldsymbol{w}_i \boldsymbol{w}_i^*, \boldsymbol{B}\rangle]_{i=1}^n$ is generally not a restricted isometry precludes our minimax analysis. However, as mentioned in Section 1.3, the restricted isometry conditions are used merely to obtain uniform accuracy guarantees. The desired uniform accuracy guarantees can be established through other approaches, as done in [24] using the *small-ball method* [30]. Moreover, if we disregard uniform guarantees and thus minimax optimality, instance accuracy guarantees of the low-rank estimation stage also can be established through the small-ball method [41] or the construction of a dual certificate [15]. Therefore, we can show that the CPR problem can be solved in our proposed framework accurately with a sample complexity that grows much slower than the ambient dimension of the target signal.

4 Numerical Experiment

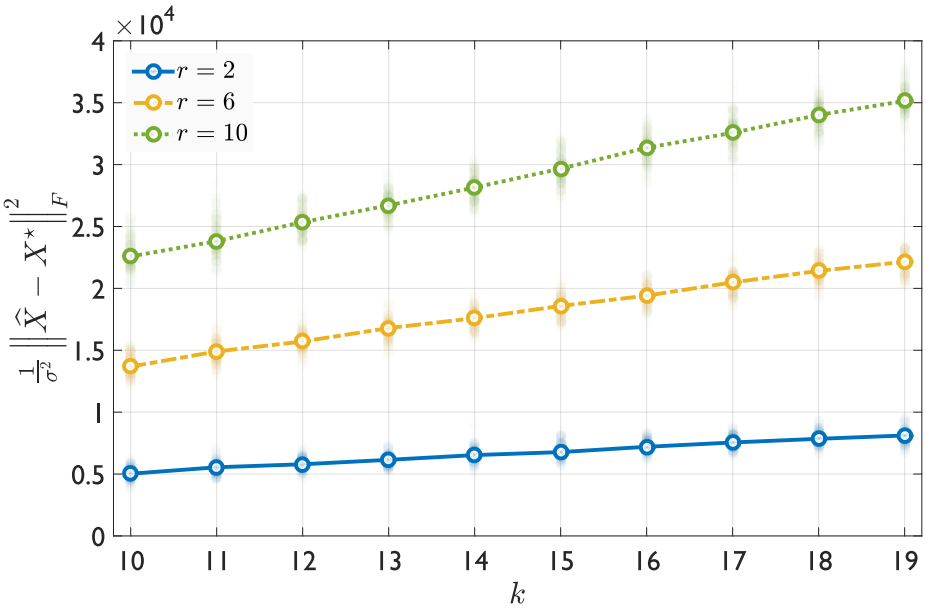
We performed a set of simulations on synthetic data in which both $\boldsymbol{\Psi}$ and $\mathcal{W}$ are considered to be Gaussian as in Corollary 1 with $m = \lceil 5k \log \frac{p_1}{k} \rceil$ and $n = 4r\,(m \vee p_2)$. We generated target matrices of size 1000×30 (i.e., $p_1 = 1000$ and $p_2 = 30$) that are factored as $\boldsymbol{X}^\star = \boldsymbol{U}\boldsymbol{V}^\mathsf{T}$ where $\boldsymbol{U} \in \mathbb{R}^{p_1\times r}$ whose k nonzero rows are chosen uniformly at random and $\boldsymbol{V} \in \mathbb{R}^{p_2\times r}$. The nonzero entries of both $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent draws from the standard Gaussian distribution. The noise variance is fixed at $\sigma^2 = 10^{-4}$. The rank of the target matrix is selected from the range $1 \leq r \leq 10$. Similarly, the row-wise sparsity is selected from the range $10 \leq k \leq 19$. For each pair of (r, k) the algorithm is tested for 100 trials. We used the TFOCS package [6] for the low-rank estimation stage (4). We also used a variant of the Alternating Direction Method of Multipliers (ADMM) adapted from [45] for the sparse estimation stage (5). Figure 1 illustrates the variation of the empirical median of $\frac{1}{\sigma^2}\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F^2$ (i.e., the normalized squared reconstruction error) versus k and r. The spread around each data point indicates the distribution of the normalized squared reconstruction error. As can be seen from the figure, $\frac{1}{\sigma^2}\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F^2$ grows almost linearly with respect to r and k if one of them is fixed which is in agreement with the theoretical results.

We also ran a simulation for the CPR described in Section 3.3 on a two-dimensional image of Saturn.[1] The target image has size $1,280 \times 1,024$, thus the ambient dimension is $p = 1,280 \cdot 1,024 = 1,310,720$. The target is *compressible* in the wavelet domain. In particular, with eight levels of

[1] Adapted from NASA's Voyager 2 image, 1981-08-24, NASA catalog #PIA01364

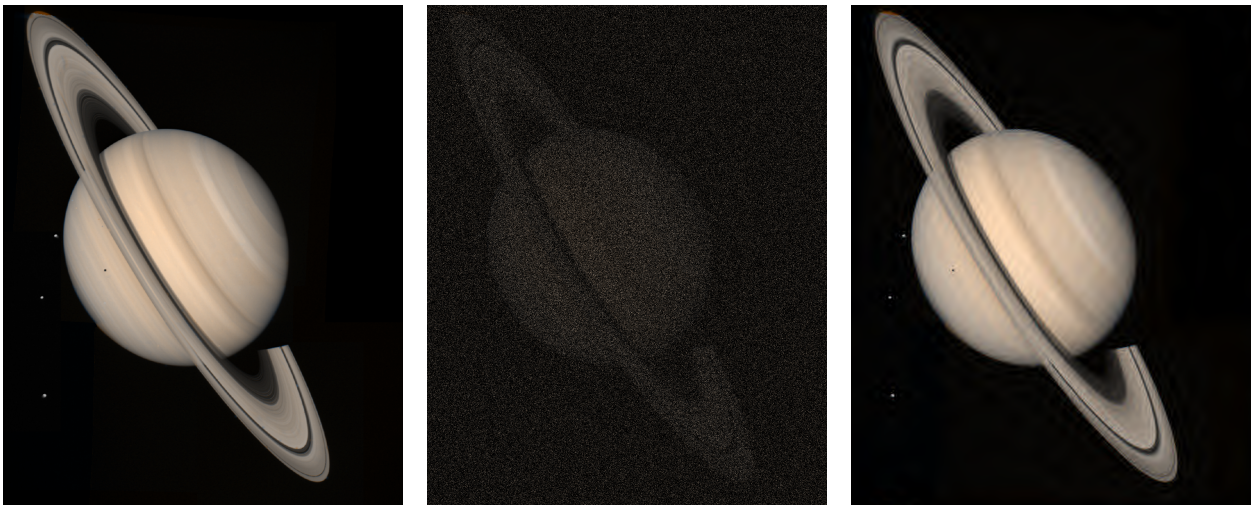


(a) $\frac{1}{\sigma^2}\left\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^\star\right\|_F^2$ vs. r for $k \in \{10, 14, 18\}$



(b) $\frac{1}{\sigma^2}\left\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^\star\right\|_F^2$ vs. k for $r \in \{2, 6, 10\}$

Figure 1: Normalized squared reconstruction error for different values of rank (i.e., r) and row-wise sparsity (i.e., k)



(a) original (b) non-sparse phase retrieval (c) sparse phase retrieval

Figure 2: Compressive phase retrieval of Saturn from nested measurements. While ordinary phase retrieval yields a poor estimate, the proposed two-stage approach that exploits sparsity has produced an estimate with relative error less than 5%.

two-dimensional 30-tap Coiflet wavelets [18] we considered the sparsity of about $k = 3,000$. The goal is to estimate the sparse/compressible wavelet coefficients of the target from the phaseless measurements of the form $\boldsymbol{y} = |\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{x}|^2$ where the modulus is taken entrywise. Here $\boldsymbol{x}$ represents the vectorized target image, $\boldsymbol{\Psi}$ effectively performs a Rademacher modulation followed by $m = \lceil 2k(1 + \log \frac{p}{k}) \rceil = 42,479$ random Fourier measurement, and $\boldsymbol{W}$ makes Fourier measurements at m randomly chosen frequencies for each of $C = 20$ masks with *octanary* patterns described in [12]. Therefore, the matrix $\boldsymbol{\Psi}$ has $m = 42,479$ rows, and the total number of measurements (i.e., the number of rows in $\boldsymbol{W}$) is $n = Cm = 849,580$. We compared the performance of our two-stage approach with the estimate produced through ordinary phase retrieval. In both cases, we used 500 iterations of the Wirtinger-Flow algorithm [12] as the phase retrieval solver. For the sparse recovery stage of our proposed method we relied on 100 iterations of the IHT algorithm [8]. Figure 2 shows, respectively from left to right, the original image, the recovered image using ordinary phase retrieval, and the recovered image through the proposed two-stage CPR. As counting the degrees of freedom would suggest, the ordinary phase retrieval is expected to fail given that we only obtained $n \approx 0.65p$ measurement. The two-stage recovery, however, produced an estimate with relative error of less than 5%. The specified algorithms and measurement schemes allowed us run the simulation without significant memory or computational requirements. Without computational and memory restrictions, the two-stage method can achieve a similar accuracy at even lower sampling rates (i.e., n/p) by using generic (e.g., Gaussian) measurements.

Funding

This work was supported by ONR grant N00014-11-1-0459 and NSF grant CCF-1422540.

References

[1] Amini, A. A. & Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, **37**(5B), 2877–2921.

[2] Bahmani, S. & Romberg, J. (2015a) Efficient Compressive Phase Retrieval with Constrained Sensing Vectors. in *Advances in Neural Information Processing Systems 28 (NIPS'15)*, pp. 523–531. Curran Associates, Inc.

[3] ——— (2015b) Lifting for blind deconvolution in random mask imaging: identifiability and convex relaxation. *SIAM Journal on Imaging Sciences*, **8**(4), 2203–2238.

[4] ——— (2015c) Sketching for Simultaneously Sparse and Low-Rank Covariance Matrices. in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'15), IEEE 6th International Workshop on*, pp. 357–360, Cancun, Mexico.

[5] Baraniuk, R., Davenport, M., DeVore, R. & Wakin, M. (2008) A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, **28**(3), 253–263.

[6] Becker, S. R., Candès, E. J. & Grant, M. C. (2011) Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, **3**(3), 165–218.

[7] Berthet, Q. & Rigollet, P. (2013) Complexity theoretic lower bounds for sparse principal component detection. in *Journal of Machine Learning Research W&CP*, vol. 30 of *Proceedings of the 26th Conference on Learning Theory (COLT'13)*, pp. 1046–1066.

[8] Blumensath, T. & Davies, M. E. (2009) Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, **27**(3), 265–274.

[9] Buja, A., Ma, Z. & Yang, D. (2013) Optimal denoising of simultaneously sparse and low rank matrices in high dimensions. in *Communication, Control, and Computing (Allerton'13), 51st Annual Allerton Conference on*, pp. 445–447. extended arXiv preprint `arXiv:1405.0338`.

[10] Cai, T., Ma, Z. & Wu, Y. (2015) Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, **161**(3-4), 781–815.

[11] Cai, T. T., Ma, Z. & Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Annals of Statistics*, **41**(6), 3074–3110.

[12] Candès, E. J., Li, X. & Soltanolkotabi, M. (2015) Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *Information Theory, IEEE Transactions on*, **61**(4), 1985–2007.

[13] Candès, E. J. & Plan, Y. (2011) Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *Information Theory, IEEE Transactions on*, **57**(4), 2342–2359.

[14] Candès, E. J. & Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, **9**(6), 717–772.

[15] Candès, E. J., Strohmer, T. & Voroninski, V. (2013) PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Communications on Pure and Applied Mathematics*, **66**(8), 1241–1274.

[16] Chen, Y., Chi, Y. & Goldsmith, A. (2015) Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming. *Information Theory, IEEE Transactions on*, **61**(7), 4034–4059.

[17] d'Aspremont, A., El Ghaoui, L., Jordan, M. I. & Lanckriet, G. R. G. (2007) A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, **49**(3), 434–448.

[18] Daubechies, I. (1992) *Ten Lectures on Wavelets*chap. Symmetry for Compactly Supported Wavelet Bases. Society for Industrial and Applied Mathematics.

[19] Eldar, Y. C. & Mishali, M. (2009) Robust Recovery of Signals From a Structured Union of Subspaces. *Information Theory, IEEE Transactions on*, **55**(11), 5302–5316.

[20] Gross, D. (2011) Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *Information Theory, IEEE Transactions on*, **57**(3), 1548–1566.

[21] Iwen, M., Viswanathan, A. & Wang, Y. (2015) Robust sparse phase retrieval made easy. *Applied and Computational Harmonic Analysis*, In press. arXiv preprint `arXiv:1410.5295 [math.NA]`.

[22] Johnstone, I. M. & Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**(486), 682–693.

[23] Koltchinskii, V., Lounici, K., Tsybakov, A. B. et al. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, **39**(5), 2302–2329.

[24] Kueng, R., Rauhut, H. & Terstiege, U. (2015) Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, In press. arXiv preprint `arXiv:1410.6913 [cs.IT]`.

[25] Laurent, B. & Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, **28**(5), 1302–1338.

[26] Lee, K., Wu, Y. & Bresler, Y. (2013) Near Optimal Compressed Sensing of Sparse Rank-One Matrices via Sparse Power Factorization. arXiv preprint `arXiv:1312.0525`.

[27] Li, X. & Voroninski, V. (2013) Sparse Signal Recovery from Quadratic Measurements via Convex Programming. *SIAM Journal on Mathematical Analysis*, **45**(5), 3019–3033.

[28] Ma, Z. & Sun, T. (2014) Adaptive sparse reduced-rank regression. arXiv preprint `arXiv:1403.1922`.

[29] Massart, P. (2007) *Concentration inequalities and model selection. Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*, no. 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer-Verlag Berlin Heidelberg.

[30] Mendelson, S. (2014) Learning without concentration. in *Journal of Machine Learning Research W&CP*, vol. 35 of *Proceedings of the 27th Conference on Learning Theory (COLT'14)*, pp. 25–39.

[31] Moravec, M. L., Romberg, J. K. & Baraniuk, R. G. (2007) Compressive phase retrieval. in *Proceedings of SPIE Wavelets XII*, vol. 6701, pp. 670120 1–11.

[32] Netrapalli, P., Jain, P. & Sanghavi, S. (2013) Phase Retrieval using Alternating Minimization. in *Advances in Neural Information Processing Systems 26 (NIPS'13)*, ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger, pp. 2796–2804. Curran Associates, Inc.

[33] Ohlsson, H., Yang, A., Dong, R. & Sastry, S. (2012) CPRL–An Extension of Compressive Sensing to the Phase Retrieval Problem. in *Advances in Neural Information Processing Systems 25 (NIPS'12)*, pp. 1367–1375.

[34] Oymak, S., Jalali, A., Fazel, M., Eldar, Y. & Hassibi, B. (2015) Simultaneously Structured Models with Application to Sparse and Low-rank Matrices. *Information Theory, IEEE Transactions on*, **61**(5), 2886–2908.

[35] Raskutti, G., Wainwright, M. & Yu, B. (2011) Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls. *Information Theory, IEEE Transactions on*, **57**(10), 6976–6994.

[36] Recht, B., Fazel, M. & Parrilo, P. A. (2010) Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, **52**(3), 471–501.

[37] Rohde, A. & Tsybakov, A. B. (2011) Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, **39**(2), 887–930.

[38] Shechtman, Y., Beck, A. & Eldar, Y. C. (2014) GESPAR: Efficient Phase Retrieval of Sparse Signals. *Signal Processing, IEEE Transactions on*, **62**(4), 928–938.

[39] Shechtman, Y., Eldar, Y. C., Szameit, A. & Segev, M. (2011) Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics Express*, **19**(16), 14807–14822.

[40] Tang, G. & Recht, B. (2014) Convex Blind Deconvolution with Random Masks. in *Classical Optics'14*, OSA Technical Digest (online), p. CW4C.1. Optical Society of America.

[41] Tropp, J. A. (2015) *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*chap. Convex Recovery of a Structured Signal from Independent Random Linear Measurements, pp. 67–101. Springer International Publishing, Cham.

[42] Tsybakov, A. B. (2008) *Introduction to Nonparametric Estimation*, Springer Series in Statistics. Springer-Verlag New York, 1st edn.

[43] Vu, V. Q., Cho, J., Lei, J. & Rohe, K. (2013) Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA. in *Advances in Neural Information Processing Systems 26 (NIPS'13)*, pp. 2670–2678.

[44] Wang, Z., Lu, H. & Liu, H. (2014) Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. in *Advances in Neural Information Processing Systems 27 (NIPS'14)*, pp. 3383–3391.

[45] Yang, J. & Zhang, Y. (2011) Alternating Direction Algorithms for ℓ_1-Problems in Compressive Sensing. *SIAM Journal on Scientific Computing*, **33**(1), 250–278.

[46] Yapar, Ç., Pohl, V. & Boche, H. (2015) Fast compressive phase retrieval from Fourier measurements. in *Acoustics, Speech and Signal Processing (ICASSP'15), IEEE International Conference on*, pp. 3267–3271.

A Auxiliary tools

To facilitate readability of the proofs, in this section we state the most important auxiliary results in our derivations that we borrowed from the literature.

Theorem 4 (Fano-type inequality [42, Theorem 2.5])**.** *Let $d(\cdot,\cdot)$ be a metric and $D\left(\mathsf{P} \parallel \mathsf{P}'\right)$ denote the Kullback-Leibler divergence of probability measures P and P'. Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ associated with probability measures $\mathsf{P}_j := \mathsf{P}_{\theta_j}$ such that:*

1. $d(\theta_j, \theta_k) \geq 2s > 0, \qquad \forall 0 \leq j < k \leq M,$
2. *P_j is absolutely continuous with respect to P_0 for all $j = 1, 2, \ldots, M$, and*

$$\frac{1}{M}\sum_{j=1}^{M} D\left(\mathsf{P}_j \parallel \mathsf{P}_0\right) \leq \alpha \log M$$

with $0 < \alpha < \frac{1}{8}$.

Then

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathsf{P}_\theta\left(d\left(\widehat{\theta},\theta\right) \geq s\right) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

Lemma 1 (Varshamov-Gilbert bound variation [29, Lemma 4.10])**.** *Let $\{0,1\}^N$ be equipped with Hamming distance d_{H} and given $1 \leq D < N$ define $\{0,1\}_D^N = \left\{\boldsymbol{x} \in \{0,1\}^N \mid d_{\mathsf{H}}(\mathbf{0}, \boldsymbol{x}) = D\right\}$. For every $\alpha \in (0,1)$ and $\beta \in (0,1)$ such that $D \leq \alpha\beta N$, there exists some subset Θ of $\{0,1\}_D^N$ with the following properties*

$$d_{\mathsf{H}}\left(\theta,\theta'\right) > 2\left(1-\alpha\right)D \;\forall\left(\theta,\theta'\right) \in \Theta^2 \text{ with } \theta \neq \theta',$$
$$\log|\Theta| \geq \rho D \log\left(\frac{N}{D}\right),$$

where

$$\rho = \frac{\alpha}{-\log\left(\alpha\beta\right)}\left(-\log\left(\beta\right) + \beta - 1\right).$$

B Proofs

B.1 Proof of the minimax upper bound

The following lemma shows the accuracy of the first stage of the estimation procedure. The proof is analogous to the standard analysis of compressive sensing and low-rank matrix estimation based on the restricted isometry property. In particular, our derivations are similar to that of [13]. We also refer to some of the lemmas proved in [13] without repeating their statements explicitly.

Lemma 2 (Low-rank reconstruction)**.** *Suppose that $n = \mathsf{O}\left(r\left(m \vee p_2\right)\right)$ and the constant c_1 in (4) is sufficiently large to guarantee that $\boldsymbol{B}^\star$ is feasible. If the restricted isometry constant $\delta_{\bullet,4r}\left(\mathcal{W}\right) < \frac{\sqrt{2}-1}{2}$, then for some constant $c_2 > 0$ depending only on the restricted isometry constants, the pre-estimate $\widehat{\boldsymbol{B}}$ obtained from (4) obeys*

$$\left\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^\star\right\|_F \leq c_2 \sigma \sqrt{r\left(m \vee p_2\right)},$$

with high probability.

Proof. It follows from optimality of $\widehat{\boldsymbol{B}}$ and feasibility of $\boldsymbol{B}^\star$ in (4) that

$$\left\|\widehat{\boldsymbol{B}}\right\|_* \leq \left\|\boldsymbol{B}^\star\right\|_*. \tag{10}$$

Let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$ be the (compact) singular value decomposition of the rank-r matrix $\boldsymbol{B}^\star \in \mathbb{R}^{m \times p_2}$. Denote the subspace of matrices that are "supported" on $\boldsymbol{U}$ and $\boldsymbol{V}$ by

$$T_0 = \left\{\boldsymbol{B} \in \mathbb{R}^{m \times p_2} \mid \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\mathsf{T}\right)\boldsymbol{B}\left(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^\mathsf{T}\right) = \boldsymbol{0}\right\}.$$

Furthermore, let $\boldsymbol{E} = \widehat{\boldsymbol{B}} - \boldsymbol{B}^\star$ and define $\boldsymbol{E}_0$ to be the projection of $\boldsymbol{E}$ onto the subspace T_0. We have

$$\begin{aligned}
\left\|\boldsymbol{B}^\star\right\|_* \geq \left\|\widehat{\boldsymbol{B}}\right\|_* &= \left\|\boldsymbol{B}^\star + \boldsymbol{E} - \boldsymbol{E}_0 + \boldsymbol{E}_0\right\|_* \\
&\geq \left\|\boldsymbol{B}^\star + \boldsymbol{E} - \boldsymbol{E}_0\right\|_* - \left\|\boldsymbol{E}_0\right\|_* \\
&= \left\|\boldsymbol{B}^\star\right\|_* + \left\|\boldsymbol{E} - \boldsymbol{E}_0\right\|_* - \left\|\boldsymbol{E}_0\right\|_*,
\end{aligned}$$

where the first and the second line follow from (10) and the triangle inequality, and the third line follows from the fact that $\boldsymbol{B}^\star$ and $\boldsymbol{E} - \boldsymbol{E}_0$ have mutually orthogonal columnspaces and rowspaces since $\boldsymbol{E} - \boldsymbol{E}_0 \in T_0^\perp$ (see [36, Lemma 2.3]). Therefore, we deduce that

$$\left\|\boldsymbol{E} - \boldsymbol{E}_0\right\|_* \leq \left\|\boldsymbol{E}_0\right\|_* \leq \sqrt{2r}\left\|\boldsymbol{E}_0\right\|_F. \tag{11}$$

Using the definition of the measurement $\boldsymbol{y} = \mathcal{W}\left(\boldsymbol{B}^\star\right) + \boldsymbol{z}$ we can write

$$\begin{aligned}
\left\|\mathcal{W}\left(\widehat{\boldsymbol{B}}\right) - \boldsymbol{y}\right\|_2^2 = \left\|\mathcal{W}\left(\boldsymbol{E}\right) - \boldsymbol{z}\right\|_2^2 &= \left\|\mathcal{W}\left(\boldsymbol{E}\right)\right\|_2^2 - 2\left\langle\mathcal{W}\left(\boldsymbol{E}\right), \boldsymbol{z}\right\rangle + \left\|\boldsymbol{z}\right\|_2^2 \\
&\geq \left\|\mathcal{W}\left(\boldsymbol{E}\right)\right\|_2^2 - 2\left\|\boldsymbol{E}\right\|_*\left\|\mathcal{W}^*\left(\boldsymbol{z}\right)\right\| + \left\|\boldsymbol{z}\right\|_2^2.
\end{aligned}$$

It is shown in [13, Lemma 1.1] that for a sufficiently large constant $c > 0$ we have $\|\mathcal{W}^*(\boldsymbol{z})\| \leq c\sigma\sqrt{m \vee p_2}$ with probability exceeding $1 - 2e^{-c_0 n}$ where c_0 is an absolute constant. Therefore, given the feasibility of $\widehat{\boldsymbol{B}}$ and (6) we obtain

$$\|\mathcal{W}(\boldsymbol{E})\|_2^2 \leq 2c\sigma \|\boldsymbol{E}\|_* \sqrt{m \vee p_2} + 2c_1\sigma^2 r(m \vee p_2).$$

Then, using the triangle inequality $\|\boldsymbol{E}\|_* \leq \|\boldsymbol{E}_0\|_* + \|\boldsymbol{E} - \boldsymbol{E}_0\|_*$ and (11) we deduce that

$$\begin{aligned}\|\mathcal{W}(\boldsymbol{E})\|_2^2 &\leq 4c\sigma\sqrt{2r} \|\boldsymbol{E}_0\|_F \sqrt{m \vee p_2} + 2c_1\sigma^2 r(m \vee p_2) \\ &\leq 4c\sigma\sqrt{2r} \|\boldsymbol{E}\|_F \sqrt{m \vee p_2} + 2c_1\sigma^2 r(m \vee p_2). \end{aligned} \tag{12}$$

For $i = 1, 2, \ldots$, define the matrices $\boldsymbol{E}_i$ recursively as the best rank-$2r$ approximation of $\boldsymbol{E} - \sum_{j=0}^{i-1} \boldsymbol{E}_j$ until $\boldsymbol{E}_i = \boldsymbol{0}$. Clearly, we have

$$\boldsymbol{E} = \sum_{j \geq 0} \boldsymbol{E}_j,$$

and thereby

$$\mathcal{W}(\boldsymbol{E}) = \sum_{j \geq 0} \mathcal{W}(\boldsymbol{E}_j).$$

To simplify the notation we use the shorthand $\delta_{\bullet,4r} = \delta_{\bullet,4r}(\mathcal{W})$ below. It follows from [13, Lemma 3.3] that for any pair of distinct indices j and j' we have

$$\left\langle \mathcal{W}(\boldsymbol{E}_j), \mathcal{W}\left(\boldsymbol{E}_{j'}\right) \right\rangle \geq -\delta_{\bullet,4r} \|\boldsymbol{E}_j\|_F \left\|\boldsymbol{E}_{j'}\right\|_F.$$

Therefore, we can expand $\|\mathcal{W}(\boldsymbol{E})\|_2^2$ and write

$$\begin{aligned}
\|\mathcal{W}(\boldsymbol{E})\|_2^2 &= \left\| \sum_{j \geq 0} \mathcal{W}(\boldsymbol{E}_j) \right\|_2^2 \\
&= \|\mathcal{W}(\boldsymbol{E}_0) + \mathcal{W}(\boldsymbol{E}_1)\|_2^2 + \left\| \sum_{j \geq 2} \mathcal{W}(\boldsymbol{E}_j) \right\|_2^2 + 2 \sum_{j \geq 2} \langle \mathcal{W}(\boldsymbol{E}_0) + \mathcal{W}(\boldsymbol{E}_1), \mathcal{W}(\boldsymbol{E}_j) \rangle \\
&= \|\mathcal{W}(\boldsymbol{E}_0) + \mathcal{W}(\boldsymbol{E}_1)\|_2^2 + \sum_{j \geq 2} \|\mathcal{W}(\boldsymbol{E}_j)\|_2^2 \\
&+ 2 \sum_{j' > j \geq 2} \left\langle \mathcal{W}\left(\boldsymbol{E}_{j'}\right), \mathcal{W}(\boldsymbol{E}_j) \right\rangle + 2 \sum_{j \geq 2} \langle \mathcal{W}(\boldsymbol{E}_0) + \mathcal{W}(\boldsymbol{E}_1), \mathcal{W}(\boldsymbol{E}_j) \rangle \\
&\geq (1 - \delta_{\bullet,4r}) \left(\|\boldsymbol{E}_0 + \boldsymbol{E}_1\|_F^2 + \sum_{j \geq 2} \|\boldsymbol{E}_j\|_F^2 \right) \\
&- 2\delta_{\bullet,4r} \sum_{j \geq 2} (\|\boldsymbol{E}_0\|_F + \|\boldsymbol{E}_1\|_F) \|\boldsymbol{E}_j\|_F - 2\delta_{\bullet,4r} \sum_{j' > j \geq 2} \|\boldsymbol{E}_j\|_F \left\|\boldsymbol{E}_{j'}\right\|_F \\
&\geq \|\boldsymbol{E}\|_F^2 - \delta_{\bullet,4r} \left(\|\boldsymbol{E}_0 + \boldsymbol{E}_1\|_F^2 + 2(\|\boldsymbol{E}_0\|_F + \|\boldsymbol{E}_1\|_F) \sum_{j \geq 2} \|\boldsymbol{E}_j\|_F \right) \\
&- \delta_{\bullet,4r} \left(\sum_{j \geq 2} \|\boldsymbol{E}_j\|_F \right)^2.
\end{aligned} \tag{13}$$

Because $\boldsymbol{E}_0$ and $\boldsymbol{E}_1$ are orthogonal we have $\|\boldsymbol{E}_0\|_F + \|\boldsymbol{E}_1\|_F \leq \sqrt{2}\,\|\boldsymbol{E}_0 + \boldsymbol{E}_1\|_F$. Furthermore, the construction of $\boldsymbol{E}_j$ guarantees that

$$\|\boldsymbol{E}_j\|_F \leq \frac{1}{\sqrt{2r}}\,\|\boldsymbol{E}_{j-1}\|_*\,,$$

for $j \geq 2$. Since for $j \geq 1$ both the columnspaces and the rowspaces of the matrices $\boldsymbol{E}_j$ are mutually orthogonal, we can invoke [36, Lemma 2.3] and write

$$\sum_{j\geq 2} \|\boldsymbol{E}_{j-1}\|_* = \left\|\sum_{j\geq 2} \boldsymbol{E}_{j-1}\right\|_* = \|\boldsymbol{E} - \boldsymbol{E}_0\|_*.$$

Therefore, in view of (11) we obtain

$$\sum_{j\geq 2} \|\boldsymbol{E}_j\|_F \leq \frac{1}{\sqrt{2r}} \sum_{j\geq 2} \|\boldsymbol{E}_{j-1}\|_* = \frac{1}{\sqrt{2r}}\,\|\boldsymbol{E} - \boldsymbol{E}_0\|_* \leq \|\boldsymbol{E}_0\|_F \leq \|\boldsymbol{E}_0 + \boldsymbol{E}_1\|_F\,.$$

Therefore, we can simplify the bound (13) to

$$\begin{aligned}\|\mathcal{W}(\boldsymbol{E})\|_2^2 &\geq \|\boldsymbol{E}\|_F^2 - 2\left(1+\sqrt{2}\right)\delta_{\bullet,4r}\,\|\boldsymbol{E}_0 + \boldsymbol{E}_1\|_F^2 \\ &\geq \left(1 - 2\left(1+\sqrt{2}\right)\delta_{\bullet,4r}\right)\|\boldsymbol{E}\|_F^2\,. \end{aligned} \tag{14}$$

If $\delta_{\bullet,4r} < \frac{\sqrt{2}-1}{2}$, then (12) and (14) yield

$$\|\boldsymbol{E}\|_F \leq c_2\sigma\sqrt{r\,(m \vee p_2)},$$

for some constant $c_2 > 0$ that only depends on the restricted isometry constants of $\mathcal{W}$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Recall that $\boldsymbol{B}^\star = \boldsymbol{\Psi}\boldsymbol{X}^\star$. Lemma 2 guarantees that $\widehat{\boldsymbol{B}}$ obtained from (4) obeys

$$\left\|\boldsymbol{\Psi}\boldsymbol{X}^\star - \widehat{\boldsymbol{B}}\right\|_F = \left\|\boldsymbol{B}^\star - \widehat{\boldsymbol{B}}\right\|_F \leq c_2\sigma\sqrt{r\,(m \vee p_2)},$$

where c_2 is constant depending only on the restricted isometry constants of $\mathcal{W}$. Therefore, $\boldsymbol{X}^\star$ is feasible for (5). The fact that $\boldsymbol{\Psi}$ is also a restricted isometry allows us to apply results from compressive sensing of block-sparse signals (see, e.g., [19, Theorem 2 and Section VI]) and obtain

$$\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \leq C\sigma\sqrt{r\,(m \vee p_2)},$$

where $C > 0$ depends on $\delta_{\bullet,4r}(\mathcal{W})$ through c_2 and on $\delta_{2k,\bullet}(\boldsymbol{\Psi})$. □

B.2 Proof of the minimax lower bound

We follow a standard strategy to establish the minimax lower bound using information theoretic tools. The minimax probability of inaccurate estimation over $\mathbb{X}_{k,r}$ can be bounded from below by the same quantity over any subset of $\mathbb{X}_{k,r}$, particularly the finite subsets. This probability can be further relaxed to the minimax probability of error in a hypothesis testing problem defined for the same subset. Therefore, to obtain a tight lower bound on the minimax probability of error, it suffices to construct a large set of hypotheses that are difficult to distinguish based on the observations. In particular, we need to choose a finite but sufficiently large number of potential target matrices in $\mathbb{X}_{k,r}$ such that they are well-separated but they are difficult to distinguish from their noisy compressive measurements. To this end, we construct two different finite subsets $\mathbb{X}'$ and $\mathbb{X}''$ of the set of $p_1 \times p_2$ matrices of rank no more than r that have at most k nonzero rows and Frobenius norm at most ε which is denoted by

$$\mathbb{X}_{k,r,\varepsilon} \coloneqq \{\boldsymbol{X} \in \mathbb{X}_{k,r} \mid \|\boldsymbol{X}\|_F \leq \varepsilon\}.$$

Construction of these two subsets depend critically on the choice of the sets $\mathbb{S}$, $\mathbb{T}'$, and $\mathbb{T}''$ that are subsets of $\{S \subset [p_1] \,|\, |S| = k\}$, $\{\pm 1\}^{r \times p_2}$, and $\{\pm 1\}^{k \times r}$, respectively. Elements of the set $\mathbb{S}$ determine the indices of the nonzero rows, whereas the elements of $\mathbb{T}'$ and $\mathbb{T}''$ determine the value of the nonzero entries. As explained below, we rely on Lemma 1, a variant of the Varshamov-Gilbert bound established in [29], to appropriately choose $\mathbb{S}$, $\mathbb{T}'$, and $\mathbb{T}''$. Finally, using the hypothesis sets $\mathbb{X}'$ and $\mathbb{X}''$ we derive a minimax lower bound by invoking Theorem 4 which is one of the Fano-type inequalities established in [42].

Proof of Theorem 2. Given $\mathbb{S}$, $\mathbb{T}'$, and $\mathbb{T}''$ define

$$\mathbb{X}' \coloneqq \left\{ \frac{\varepsilon}{\sqrt{kp_2}} \begin{bmatrix} \boldsymbol{I}_{p_1,S} & \boldsymbol{0}_{r\lceil \frac{k}{r} \rceil - k} \end{bmatrix} \left(\boldsymbol{1}_{\lceil \frac{k}{r} \rceil \times 1} \otimes \boldsymbol{T} \right) \mid S \in \mathbb{S} \text{ and } \boldsymbol{T} \in \mathbb{T}' \right\},$$

and

$$\mathbb{X}'' \coloneqq \left\{ \frac{\varepsilon}{\sqrt{kp_2}} \boldsymbol{I}_{p_1,S} \left(\boldsymbol{T} \otimes \boldsymbol{1}_{1 \times \lceil \frac{p_2}{r} \rceil} \right) \boldsymbol{I}_{r\lceil \frac{p_2}{r} \rceil,[p_2]} \mid S \in \mathbb{S} \text{ and } \boldsymbol{T} \in \mathbb{T}'' \right\}.$$

In words, each matrix in $\mathbb{X}'$ basically consists of $\frac{k}{r}$ copies of an $r \times p_2$ binary matrix that are stacked on top of each other and interleaved with all-zero rows. Clearly, the matrices in $\mathbb{X}'$ have k nonzero rows, are of rank at most r, and have Frobenius norm of ε, showing that $\mathbb{X}' \subset \mathbb{X}_{k,r,\varepsilon}$. Similarly, each matrix in $\mathbb{X}''$ is essentially a horizontal concatenation of $\frac{p_2}{r}$ copies of an $k \times r$ binary matrix that is interleaved row-wise with all-zero rows. It is straightforward to verify that the matrices in $\mathbb{X}''$ have k nonzero rows, are of rank at most r, and have Frobenius norm of ε, which show that $\mathbb{X}'' \subset \mathbb{X}_{k,r,\varepsilon}$ as well.

We use the Varshamov-Gilbert bound as stated in Lemma 1 to choose sufficiently large sets $\mathbb{S}$, $\mathbb{T}'$, and $\mathbb{T}''$. Treating each of the sets in $\mathbb{S}$ as a binary sequence of length p_1, Lemma 1 guarantees existence of a set $\mathbb{S}$ of k-subsets of $[p_1]$ such that

$$\log |\mathbb{S}| \geq \frac{4}{25} k \log \frac{p_1}{k}, \qquad \text{and} \qquad d_{\mathsf{H}}(S_1, S_2) \geq \frac{1}{4}k, \ \forall (S_1, S_2) \in \mathbb{S}^2, \ S_1 \neq S_2.$$

We can also treat each matrix in $\mathbb{T}'$ as a binary string of length rp_2, apply Lemma 1, and show that there exists a set $\mathbb{T}'$ of matrices in $\{\pm 1\}^{r\times p_2}$ that satisfies

$$\log\left|\mathbb{T}'\right| \geq \frac{3}{25}rp_2, \qquad \text{and} \qquad d_{\mathsf{H}}\left(\boldsymbol{T}_1,\boldsymbol{T}_2\right) \geq \frac{1}{8}rp_2,\ \forall\left(\boldsymbol{T}_1,\boldsymbol{T}_2\right)\in\mathbb{T}'^2,\ \boldsymbol{T}_1\neq\boldsymbol{T}_2.$$

Similarly, there exists a set $\mathbb{T}''$ of matrices in $\{\pm 1\}^{k\times r}$ such that

$$\log\left|\mathbb{T}''\right| \geq \frac{3}{25}rk, \qquad \text{and} \qquad d_{\mathsf{H}}\left(\boldsymbol{T}_1,\boldsymbol{T}_2\right) \geq \frac{1}{8}rk,\ \forall\left(\boldsymbol{T}_1,\boldsymbol{T}_2\right)\in\mathbb{T}''^2,\ \boldsymbol{T}_1\neq\boldsymbol{T}_2.$$

Let $(S_1,\boldsymbol{T}_1)$ and $(S_2,\boldsymbol{T}_2)$ be two distinct pairs in $\mathbb{S}\times\mathbb{T}'$ using which we can construct the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in $\mathbb{X}'$, respectively. If $S_1\neq S_2$, then counting only the rows of $\boldsymbol{X}_1-\boldsymbol{X}_2$ that are not in the set $S_1\cap S_2$ we obtain

$$\left\|\boldsymbol{X}_1-\boldsymbol{X}_2\right\|_F \geq \frac{\varepsilon}{\sqrt{kp_2}}\sqrt{d_{\mathsf{H}}\left(S_1,S_2\right)p_2} \geq \frac{\varepsilon}{2}.$$

Furthermore, if $S_1=S_2$ and $\boldsymbol{T}_1\neq\boldsymbol{T}_2$, then we have

$$\left\|\boldsymbol{X}_1-\boldsymbol{X}_2\right\|_F \geq \frac{2\varepsilon}{\sqrt{kp_2}}\sqrt{d_{\mathsf{H}}\left(\boldsymbol{T}_1,\boldsymbol{T}_2\right)\left\lfloor\frac{k}{r}\right\rfloor} \geq \varepsilon\sqrt{\frac{r\left\lfloor\frac{k}{r}\right\rfloor}{2k}} \geq \frac{\varepsilon}{2}.$$

Therefore, the above inequalities and the fact that $\log|\mathbb{X}'|=\log|\mathbb{S}|+\log|\mathbb{T}'|$ show that the set $\mathbb{X}'$ obeys

$$\log\left|\mathbb{X}'\right| \geq \frac{4}{25}k\log\frac{p_1}{k}+\frac{3}{25}rp_2, \quad \text{and} \quad \left\|\boldsymbol{X}_1-\boldsymbol{X}_2\right\|_F \geq \frac{\varepsilon}{2},\ \forall\left(\boldsymbol{X}_1,\boldsymbol{X}_2\right)\in\mathbb{X}'^2,\ \boldsymbol{X}_1\neq\boldsymbol{X}_2 \tag{15}$$

Similarly, we can show that $\mathbb{X}''$ obeys

$$\log\left|\mathbb{X}''\right| \geq \frac{4}{25}k\log\frac{p_1}{k}+\frac{3}{25}rk, \quad \text{and} \quad \left\|\boldsymbol{X}_1-\boldsymbol{X}_2\right\|_F \geq \frac{\varepsilon}{2},\ \forall\left(\boldsymbol{X}_1,\boldsymbol{X}_2\right)\in\mathbb{X}''^2,\ \boldsymbol{X}_1\neq\boldsymbol{X}_2 \tag{16}$$

For any matrix $\boldsymbol{X}\in\mathbb{R}^{p_1\times p_2}$ let $\mathsf{P}_{\boldsymbol{X}}$ denote the Gaussian distribution $\mathsf{N}\left(\mathcal{A}\left(\boldsymbol{X}\right),\sigma^2\boldsymbol{I}\right)$ which is the distribution of the measurement $\boldsymbol{y}$ if $\boldsymbol{X}$ is the target matrix. Recall that by assumption $\mathcal{A}$ is a restricted isometry over $\mathbb{X}_{k,r}$ as defined by (3) with the restricted isometry constant $\delta=\delta_{k,r}\left(\mathcal{A}\right)$. For any $\boldsymbol{X}\in\mathbb{X}'\subset\mathbb{X}_{k,r}$ the KL-divergence between $\mathsf{P}_{\boldsymbol{X}}$ and $\mathsf{P}_{\boldsymbol{0}}$ can be bounded as

$$D\left(\mathsf{P}_{\boldsymbol{X}}\|\mathsf{P}_{\boldsymbol{0}}\right)=\frac{\left\|\mathcal{A}\left(\boldsymbol{X}\right)\right\|_2^2}{2\sigma^2}\leq\frac{\gamma_{k,r}\left\|\boldsymbol{X}\right\|_F^2}{2\sigma^2}=\frac{\gamma_{k,r}\varepsilon^2}{2\sigma^2},$$

where the inequality follows from restricted isometry assumption on $\mathcal{A}$. Therefore, we have

$$\frac{1}{|\mathbb{X}'|}\sum_{\boldsymbol{X}\in\mathbb{X}'}D\left(\mathsf{P}_{\boldsymbol{X}}\|\mathsf{P}_{\boldsymbol{0}}\right)\leq\frac{\gamma_{k,r}\varepsilon^2}{2\sigma^2}.$$

Suppose that we have

$$\frac{\gamma_{k,r}\varepsilon^2}{2\sigma^2}\leq\alpha'\log\left|\mathbb{X}'\right|,$$

holds for some $\alpha' \in \left(0, \frac{1}{8}\right)$. Then, because any matrix $\boldsymbol{X} \in \mathbb{X}'$ also satisfies $\|\boldsymbol{X} - \boldsymbol{0}\|_F = \|\boldsymbol{X}\|_F = \varepsilon > \frac{\varepsilon}{2}$, we can use (15) and invoke Theorem 4 to guarantee that

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r,\varepsilon}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq \frac{\varepsilon}{4}\right) \geq \frac{\sqrt{|\mathbb{X}'|}}{1+\sqrt{|\mathbb{X}'|}} \left(1 - 2\alpha' - \sqrt{\frac{2\alpha'}{\log|\mathbb{X}'|}}\right).$$

From (15) we have the crude lower bound $\log|\mathbb{X}'| \geq \frac{3}{25}$. Therefore, with $\alpha' = 5 \times 10^{-5}$ we can choose

$$\varepsilon = 10^{-2}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + rp_2}{\gamma_{k,r}}},$$

and obtain

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r,\varepsilon}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq 2.5 \times 10^{-3}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + rp_2}{\gamma_{k,r}}}\right) > \frac{1}{2}. \tag{17}$$

Furthermore, for the set $\mathbb{X}''$ if we have

$$\frac{(1+\delta)\,\varepsilon^2}{2\sigma^2} \leq \alpha''\log\left|\mathbb{X}''\right|,$$

for some $\alpha'' \in \left(0, \frac{1}{8}\right)$, then using (16) we can apply Theorem 4 and show that

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r,\varepsilon}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq \frac{\varepsilon}{4}\right) \geq \frac{\sqrt{|\mathbb{X}''|}}{1+\sqrt{|\mathbb{X}''|}} \left(1 - 2\alpha'' - \sqrt{\frac{2\alpha''}{\log|\mathbb{X}''|}}\right).$$

Similar to the argument for $\mathbb{X}'$, with $\alpha'' = 5 \times 10^{-5}$ we can choose

$$\varepsilon = 10^{-2}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + rk}{\gamma_{k,r}}},$$

and obtain

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r,\varepsilon}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq 2.5 \times 10^{-3}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + rk}{\gamma_{k,r}}}\right) > \frac{1}{2}. \tag{18}$$

Then with $\varepsilon = 2.5 \times 10^{-3}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + r(k\vee p_2)}{\gamma_{k,r}}}$ we can deduce from (17) and (18) that

$$\inf_{\widehat{\boldsymbol{X}}} \sup_{\boldsymbol{X}^\star \in \mathbb{X}_{k,r,\varepsilon}} \mathsf{P}\left(\left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\right\|_F \geq 2.5 \times 10^{-3}\sigma\sqrt{\frac{k\log\frac{p_1}{k} + r\,(k\vee p_2)}{\gamma_{k,r}}}\right) > \frac{1}{2},$$

which also guarantees the desired minimax lower bound on $\mathbb{X}_{k,r}$. □